基于改进的深度神经网络的人体动作识别模型 *

何冰倩，魏　维，张　斌，高联欣，宋岩贝

(成都信息工程大学 计算机学院, 成都 610225)

摘　要：针对现有人体动作识别方法需输入固定长度的视频段、未充分利用时空信息等问题，提出一种基于时空金字塔和注意力机制相结合的深度神经网络模型，将包含时空金字塔的3D-CNN和添加时空注意力机制的LSTM模型相结合，实现了对视频段的多尺度处理和对动作的复杂时空信息的充分利用。以RGB图像和光流场作为空域和时域的输入，以融合金字塔池化层的运动和外观特征后的融合特征作为融合域的输入，最后采用决策融合策略获得最终动作识别结果。在UCF101和HMDB51数据集上进行实验，分别取得了94.2%和70.5%的识别准确率。实验结果表明，改进的网络模型在基于视频的人体动作识别任务上获得了较高的识别准确率。

关键词：动作识别；深度学习；时空金字塔；注意力机制；卷积神经网络

中图分类号：TP391.41　　doi: 10.3969/j.issn.1001-3695.2018.06.0361

Improved deep convolutional neural network for human action recognition

He Bingqian, Wei Wei, Zhang Bin, Gao Lianxin, Song Yanbei

(*College of Computer Science & Technology*, *Chengdu University of Information Technology*, *Chengdu* 610225, *China*)

Abstract: Aiming at the problem that the existing human motion recognition method needs to input a fixed length video segment and underutilize the spatiotemporal information, this paper proposed a deep neural network model based on the combination of space-time pyramid and attention mechanism. This improved architecture combined 3D-CNN including spatiotemporal pyramids with LSTM model with spatio-temporal attention mechanism, and realized multi-scale processing of video segments and full utilization of complex spatio-temporal information of actions. For the architecture, the inputs of spatial and temporal domain were RGB image and the optical flow, the input of the fusion domain was the fusion feature of the motion and appearance features of the pyramid pooling layer. Finally, the final motion recognition result was obtained through the decision fusion strategy. Experiments were performed on the UCF101 and HMDB51 datasets, achieving 94.2% and 70.5% recognition accuracy, respectively. The experimental results show that the improved network model achieves high recognition accuracy in video-based human motion recognition tasks.

Key words: action recognition; deep learning; spatiotemporal pyramid; attention module; convolutional neural network

0　引言

人体行为识别在机器人交互、虚拟现实、家庭和公共安全等领域的广泛应用，使其正逐渐成为计算机视觉最活跃的研究领域之一。目前的识别算法和模型可以大概地分为两类，一类是基于传统手选特征的识别算法[1~5]，一类是基于深度学习的识别算法和模型[6~13]。其中，基于深度学习的方法在各类具有挑战性的视频数据集上展现出了优于传统方法的较大优势。尽管如此，如何准确地区分不同类别的行为动作仍然存在巨大的挑战性。比如光照或遮挡等视频环境因素、动作类别的类间和类内差异、视频数据集较少，这些问题都对鲁棒特征提取和动作分类构成了巨大挑战。

为了突破卷积神经网络只应用于二维图像这一局限并且能够有效地将视频分析中的运动信息结合起来，文献[14]提出在CNN卷积层中执行三维卷积，从而捕获空间和时间维度的区分性特征，但是该模型仍然不能充分利用视频的时空特征。文献[6]为了更好地利用视频数据中的时间信息，提出了结合空间域和时间域的双流卷积网络（two-stream convolutional networks），两个卷积网络分别以视频数据的 RGB 图像和视频帧的光流作为输入，然后提取动作表示的视频帧的时间和空间特征，最后通过融合分类识别，该模型在一定程度上利用了视频序列的时空特征，但是由于只关注了当前步骤的卷积映射，可能不足以

收稿日期：2018-06-21；**修回日期：**2018-08-22　　**基金项目：**四川省教育厅重点科研项目（17ZA0064）

作者简介：何冰倩（1994-），女，四川阆中人，硕士研究生，主要研究方向为图形图像处理（dandelionqian@foxmail.com）；魏维（1976-），男，教授，博士，主要研究方向为图形图像处理；张斌（1992-），男，硕士研究生，主要研究方向为图形图像处理；高联欣（1994-），男，硕士研究生，主要研究方向为图形图像处理；宋岩贝（1994-），男，硕士研究生，主要研究方向为图形图像处理.

捕获不同类别动作的复杂时空线索[13]。目前基于 CNN 的识别模型都仅仅是捕获了短时间规模的时空特征，无法表示长时间的变化。经过一些文献[9,10,15,16]的实验证明，循环神经网络（recurrent neural networks, RNN）能在一定程度上解决这个问题，尤其是对视频序列能够较好有效建模的长时短期记忆模型（long short-term memory，LSTM）[17]。但是，在这些模型中，LSTM 的输入是直接从 CNN 的全连接层中提取的高级特征，而这些特征缺乏时空特征细节。

针对上述问题，本文在时空双流卷积网络识别模型的基础上，提出了一种结合了时空金字塔和注意力机制的深度神经网络模型（deep neural network combining spatial-temporal pyramid and attention mechanism，STPP and attention-mechanism network）。本文模型首先将视频序列的 RGB 图像和视频帧的光流分别通过 3D 卷积神经网络获取时空卷积特征映射，然后利用时空金字塔池化（spatial temporal pyramid pooling，STPP）来聚合局部时空信息形成固定长度的特征向量，再通过时空特征融合策略在 STPP 层对时空特征进行有效融合，最后将时空 3D 双流网络提取到的时空特征和融合后的特征分别输入到具有时空注意力机制的 LSTM 模型和普通 LSTM 模型中进行建模，对模型分类结果进行融合从而获得最终的人体动作分类结果。本文在数据集 UCF101 和 HMDB51 上进行人体动作识别实验，实验结果表明本文提出的基于结合时空金字塔和注意力机制的深度神经网络模型能够有效识别视频中的人体动作。

1 相关工作

深度学习在计算机视觉的图像识别领域取得的好成绩使得深度学习的方法，尤其是卷积神经网络，在计算机视觉领域得到了广泛的研究和应用。相对于静态图像而言，视频序列不仅具有外观信息还具有运动信息[18]，因此，最近的一些研究开始尝试设计能够有效利用视频序列外观和运动信息的基于卷积神经网络的动作识别模型。文献[19]研究比较了多种 CNN 的连接方式中的三种广泛使用的方法，即后期融合、早期融合和慢速融合。实验结果说明这些方法都不能充分利用运动信息，只能对单个框架进行适度的改进。文献[20]在 UCF101 和 sports-1M 上训练了更深的 CNN 模型，称为 C3D 网络模型。该模型近似于一个 3D 版本的 VGGnet 模型[7]，包含了一个 3D 卷积滤波器和一个同时对时间域和空间域进行操作的 3D 池化层。文献[6]提出的双流卷积神经网络，通过对视频帧的光流训练第二个 CNN 流，一定程度上弥补了叠加的 RGB 流不能充分利用时间信息的缺陷，为动作识别方法带来了一定的性能增益。该模型也被广泛用于许多其他动作识别方法[8,9,21,22]。

但原始的双流卷积神经网络模型有两个主要问题：a)该模型由于只包含 10 个连续的光流帧而不能捕获长期的时间信息；b)该模型是对空间域和时间域分别进行训练，最终预测是根据两个分类器的输出平均而得到的，因此不能有效地学习时间流和空间流之间的时空关系。对于这些问题，文献[10]提出了一种基于 LSTM 的动作识别分类方法，以此来融合更长期的视频序列中的特征。文献[23]提出了通过具有稀疏采样的分段网络架构来模拟长期时间结构。文献[24]通过研究在时间和空间上组合网络的多种方式，提出了一种时空融合方法，并且认为双流网络应该在最后的卷积层进行融合。尽管上述文献的方法或模型对原始双流卷积神经网络存在的问题进行了一定的改善，但是仍然存在丢失重要的时空线索的问题，使得模型不能获取充分的人体动作的时空关系，以及不能对任意长度的视频段进行特征提取的问题，大都需要对视频段进行手动的预处理。

基于对上述问题的考虑，本文在文献[24]的基础上，提出了一种基于结合时空金字塔和注意力机制的深度神经网络的人体动作识别模型。对于需要直接处理任意长度的视频段的任务，本文对原 C3D 网络进行简单改进，在最后一个卷积层后加入时空金字塔池化，使得模型能够生成固定长度的特征向量。同时由于时空金字塔是从多角度对特征映射进行处理，使得模型能够得到更深层的特征表示，从而提高识别精度。对于捕捉人体动作之间复杂的时空线索的任务，本文设计了添加时空注意力机制的 LSTM 模型，该模型不仅能捕获长期的时间信息，还能通过时空注意力机制捕获人体动作的复杂时空线索。本文还在在模型中添加了时空特征融合模块，使得模型尽量不丢失重要的动作特征。

2 结合时空金字塔和注意力机制的深度神经网络人体动作识别模型设计

2.1 整体框架

本文提出的网络框架图如图 1 所示。该模型主要包含三个模块：结合时空金字塔池化的时空双流三维卷积神经网络；空间与时间域的特征融合；包含时空注意力机制的长期短时记忆模型。

对于第一个模块，本文采用文献[6]的时空双流模型和文献[20]中的 C3D 网络结构，并对其进行改进然后形成本文模型中的时空双流三维卷积神经网络模块。时间流和空间流深度卷积神经网络网络都由 5 组卷积层、4 个最大池化、1 个时空金字塔池化和 2 个全连接层组成，即将原来 C3D 网络的最后一个最大池化层改为时空金字塔池化。时空金字塔池化不仅能解决输入尺寸不一的情况，还能通过不同角度的特征提取方法提取出更深的特征，从而提高识别精度。具体来说，从 1 到 5 的 5 组卷积层的过滤器数目分别是 64、128、256、512、512，2 个全连接层是 4096 个单元。根据文献[19]对卷积层的不同深度的内核实验研究结果，3x3x3 的核尺寸大小是对所有卷积层来说最佳的选择，因此，在此模块中，所有卷积层均采用 3x3x3 的内核大小，步长为 1x1x1。对于最大池化层，除了第一个最大池化的核大小是 2x2x1，其余 3 个最大池化层的核大小为 2x2x2。第一个模块直接连接到第三个模块。第二个模块主要是对时空双流提取到的特征进行融合，然后连接到第三个模块中不包含时空注意力机制的 LSTM 模型。该模块在第一模块的 STPP 层进

行。第三个模块是添加了注意力机制的 LSTM 模型。LSTM 模型本身作为循环神经网络，能够通过保存时间序列信息来捕获长期的时空依赖关系，还能有效避免梯度消失现象，而该模块较于原始的 LSTM 模型还能够捕获更复杂的时空线索，从而提高识别准确率。总体而言，本文的网络框架包含了特征级的数据融合和决策级的融合，通过这两种层面的融合方法使得该网络模型对人体动作的识别更加准确。

本文模型在 ImageNet 上进行预训练和微调后，将视频序列的 RGB 图像数据和视频帧的光流数据输入到该模型中，通过训练两个三维卷积神经网络来提取时间流和空间流特征，再利用时空金字塔来提取固定长度的特征向量，然后通过两个全连接层提取视频帧的深度特征。同时利用时空特征融合策略融合从 STPP 层中提取到的人体动作深层特征，最后通过包含时空注意力机制的 LSTM 模型对时空特征进行建模，进而获得分类结果。

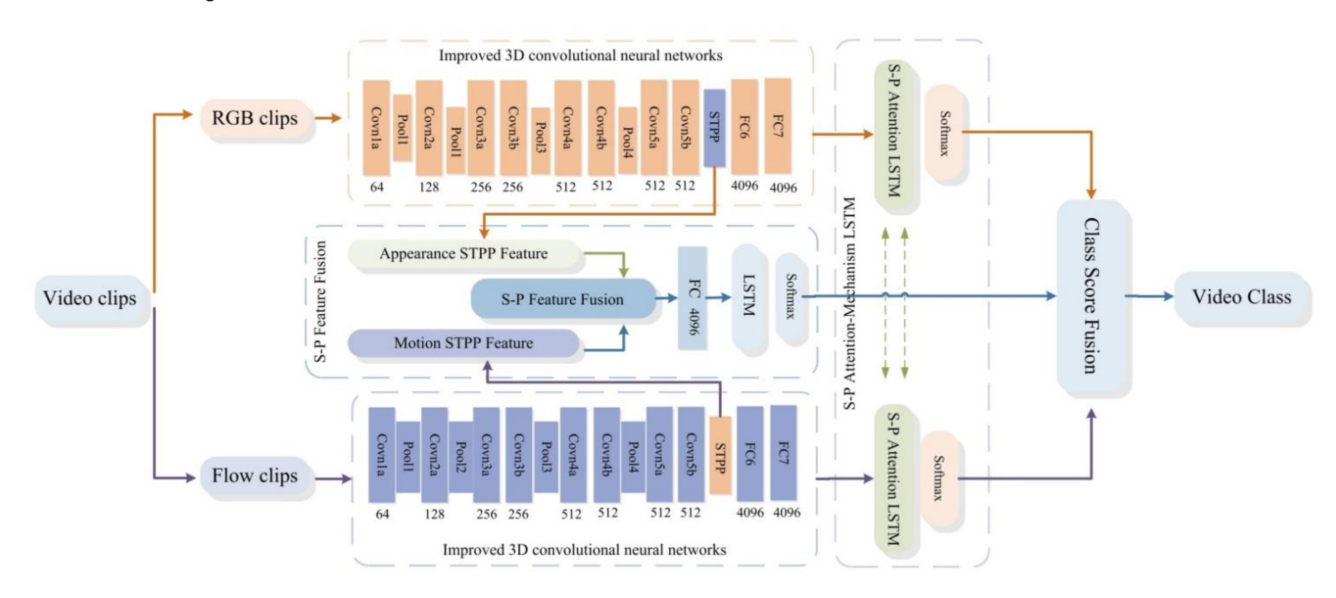


图 1　结合时空金字塔和注意力机制的深度神经网络的人体动作识别模型

2.2　时空金字塔池化

为了对任意大小长度的视频序列都能采用本文模型进行处理，本文利用时空金字塔池化（STPP）来生成固定长度的特征向量。同时，由于时空金字塔池化层能从不同角度对卷积得到的特征映射进行特征提取，能一定程度上为人体动作识别提高精度。

在该层中，可以输入任意大小和长度的视频序列。假定输入的 RGB 和光流图片序列的大小为 $l\times h\times w$，而最后一层卷积的特征映射大小为 $T\times H\times W$，其中 l 是长度（帧数），T 为池化立方体的时间大小，h、H 和 w、W 是帧的高度和宽度。本文将对输入到 STPP 的每个时空立方体的响应值和最大化操作集中到一起。不同于文献[20]中一般滑动窗口的池化操作，STPP 层的滑动窗口大小是在给定池化水平内动态调整的。简单来说，将 $P(p_t, p_s)$ 作为时空池化水平。那么，每个立方体大小为 $T/p_t \times H/p_s \times W/p_s$，其中，$p_t$ 是时间池化水平，p_s 是空间池化水平。由于每段视频序列的时间尺度比对应的空间尺度小，本文将 p_t 的值设为 1。当 $p_s=4,2,1; p_t=1$，每个输入的视频片段会生成固定长度的描述符，从而 STPP 通过聚合局部时空信息形成固定长度的特征向量。

2.3　时空特征融合

对于基于视频的人体动作识别，提取的特征不仅是静态的视觉特征，还有动态的时间运动特征。合适且效果好的特征融合方法能够利用两种特征的相关性来生成更多元的混合特征。因此，本文根据文献[24]的研究，提出了时空特征融合框架。

对于模型输入的第 t 段视频序列，可以在第一模块的 STPP 层得到两个 STPP 特征，将其表示为 x_t^a 和 x_t^m，其中，x_t^a 代表第 t 段序列的 RGB 特征，即外观特征；x_t^m 代表第 t 段序列的光流特征，即运动特征。本文采用早期融合方法（元素串联）来融合上述两个 STPP 特征，并生成一个新的融合特征 x_t^f。然后，将所得到的混合特征通过一个 4096 个单元的全连接层再链接到本文的第三模块，即利用长时短期记忆模型对融合特征进行建模以及分类。

2.4　包含时空注意力机制的 LSTM 模型

在该模块中，本文设计了包含时空注意力机制的 LSTM 模型（S-P attention-mechanism LSTM）来对前期获取的深层特征进行建模。LSTM 作为一种循环神经网络，能够通过保存时间序列信息来捕获长期的时空依赖关系，同时，不同于原始的 RNN，LSTM 在经过反向传播训练后不会出现梯度消失的情况。用于人体动作识别的视频序列往往包含很多时空线索，如果直接将第一模块的全连接层的特征输入到 LSTM 中，模型将会不足以捕捉不同动作的复杂时空线索，因此，为了能够进一步捕捉到有用的特征，本文在基础的 LSTM 模型中加入了时空注意力机制。

LSTM 的一个单元如图 2 所示，图示中*代表 a 或者 m。本文将第一模块全连接层得到的高维特征描述为 X_t^a 和 X_t^m，分别

表示第 t 段视频序列的外观和运动特征；将第二模块全连接层得到的融合特征描述为 X_t^f。X_t^* 作为 S-P Attention LSTM 模块的输入。

i_t^*、f_t^*、o_t^* 分别代表输入门、遗忘门和输出门，g_t^*、c_t^*、h_t^*、Y_t^* 分别代表记忆调制状态、内核状态（记忆状态）、隐藏状态和输出。对于融合特征 X_t^f，本文将其输入到普通 LSTM 中，其实现公式如下：

$$i_t^f = \sigma_f\left(w_{xi}^f X_t^f + w_{hi}^f h_{t-1}^f + b_i^f\right) \tag{1}$$

$$f_t^f = \sigma_f\left(w_{xf}^f X_t^f + w_{hf}^f h_{t-1}^f + b_f^f\right) \tag{2}$$

$$o_t^f = \sigma_f\left(w_{xo}^f X_t^f + w_{ho}^f h_{t-1}^f + b_o^f\right) \tag{3}$$

$$g_t^f = tanh\left(w_{xg}^f X_t^f + w_{hg}^f h_{t-1}^f + b_g^f\right) \tag{4}$$

$$c_t^f = f_t^f \odot c_{t-1}^f + i_t^f \odot g_t^f \tag{5}$$

$$h_t^f = o_t^f \odot \tanh\left(c_t^f\right) \tag{6}$$

其中，h_{t-1}^f 表示前一个隐藏状态，w_{x*}^f 和 w_{h*}^f 分别是输入向量和隐藏状态的权重矩阵，b_*^f 代表偏差向量，$\sigma(\cdot)$ 和 $\tanh(\cdot)$ 分别表示激活函数中的 *sigmoid* 函数和 *tanh* 函数，$\odot$ 表示哈达玛积，即矩阵元素对应相乘。

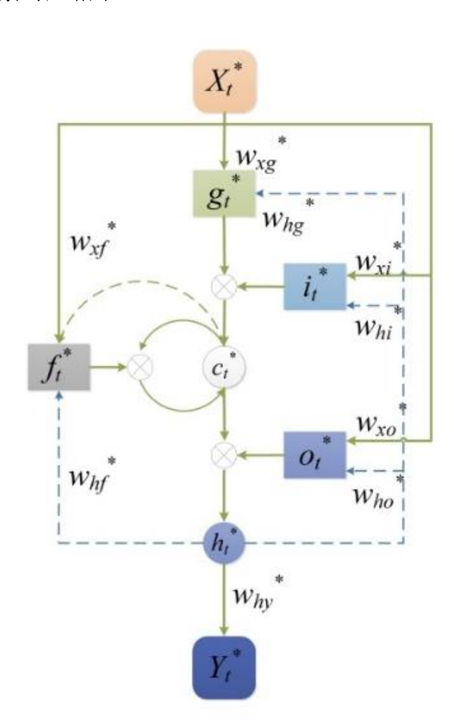


图 2　普通 LSTM 模型的一个单元

2.5　LSTM 的时空注意力机制

LSTM 的时空注意力机制模型如图 3 所示，时空注意力机制同时作用于空间域和时间域。空间域的输入为 X_t^a，时间域的输入为 X_t^m，为防止描述重复，将该模块的输入统一表示为 X_t^*，其中*代表 a 或者 m。为了找到第 t 段视频序列中具有重要描述意义的特征向量，本文对每个流先进行空间注意力运算，其计算过程如下。

以 LSTM 单元的前一个隐藏状态 h_{t-1}^* 为例，首先利用式(7)(8)来计算第 t 段视频序列中第 k 个特征向量对第 n 个特征向量的空间注意力概率 $\alpha_t^*(n,k)$：

$$\tilde{\alpha}_t^*(n,k) = \mu_\alpha^* \tanh\left(A_h^* h_{t-1}^* + A_X^* X_t^*(n,k) + b_\alpha^*\right) \tag{7}$$

$$\alpha_t^*(n,k) = \frac{exp\left(w_t^\alpha \tilde{\alpha}_t^*(n,k)\right)}{\sum_{l=1}^{L} exp\left(w_t^\alpha \tilde{\alpha}_t^*(n,l)\right)} \tag{8}$$

其中：μ_α^*、A_h^*、A_X^*、w_t^α 是空间注意力机制的权值矩阵，b_α^* 是空间注意力机制的偏置向量，L 是第 t 段视频序列中的帧数目，$\tilde{\alpha}_t^*$ 是未规范化的注意力概率。然后，利用式(9)获取第 n 个特征向量的空间特征向量：

$$L_t^*(n) = \sum_{k=1}^{L} \alpha_t^*(n,k) X_t^*(n,k) \quad n = 1,\cdots,T \tag{9}$$

在得到具有空间重要性的空间特征向量 $L_t^*(n)$ 后，本文对其进行时间注意力计算，同空间注意力计算类似，先计算时间注意力概率 $\beta_t^*(n)$，计算公式如下：

$$\tilde{\beta}_t^*(n) = \mu_\beta^* tanh\left(B_h^* h_{t-1}^* + B_X^* L_t^*(n) + b_\beta^*\right) \tag{10}$$

$$\beta_t^*(n) = \frac{exp\left(w_t^\beta \tilde{\beta}_t^*(n)\right)}{\sum_{j=1}^{T} exp\left(w_t^\beta \tilde{\beta}_t^*(j)\right)} \tag{11}$$

其中：μ_β^*、B_h^*、B_X^*、w_t^β 是时间注意力机制的权值矩阵，b_β^* 是时间注意力机制的偏置向量，T 是第 t 段视频序列的总特征向量数。$\beta_t^*(n)$ 反映了第 n 个特征向量对第 t 段视频序列的时间重要性。根据式(12)计算最后时空注意力捕捉到的重要的时空特征 Φ_t^*：

$$\Phi_t^* = \sum_{n=1}^{T} \beta_t^*(n) L_t^*(n) \tag{12}$$

由于此时得到的上下文特征 Φ_t^* 与当前步骤的预测是紧密相关的，本文将其作为 LSTM 模型除了原本特征向量 X_t^* 之外的额外输入，具体计算公式如下，

$$i_t^* = \sigma_*\left(w_{Xi}^* X_t^* + w_{\Phi i}^* \Phi_t^* + w_{hi}^* h_{t-1}^* + b_i^*\right) \tag{13}$$

$$f_t^* = \sigma_*\left(w_{Xf}^* X_t^* + w_{\Phi f}^* \Phi_t^* + w_{hf}^* h_{t-1}^* + b_f^*\right) \tag{14}$$

$$o_t^* = \sigma_*\left(w_{Xo}^* X_t^* + w_{\Phi o}^* \Phi_t^* + w_{ho}^* h_{t-1}^* + b_o^*\right) \tag{15}$$

$$g_t^* = tanh\left(w_{Xg}^* X_t^* + w_{\Phi g}^* \Phi_t^* + w_{hg}^* h_{t-1}^* + b_{fg}^*\right) \tag{16}$$

$$c_t^* = f_t^* \odot c_{t-1}^* + i_t^* \odot g_t^* \tag{17}$$

$$h_t^* = o_t^* \odot \tanh\left(c_t^*\right) \tag{18}$$

其中：w 是 LSTM 模型中的权值矩阵，b 是偏置向量，$\sigma(\cdot)$ 和 $\tanh(\cdot)$ 分别表示激活函数中的 *sigmoid* 函数和 *tanh* 函数，$\odot$ 表示哈达玛积，即矩阵元素对应相乘。

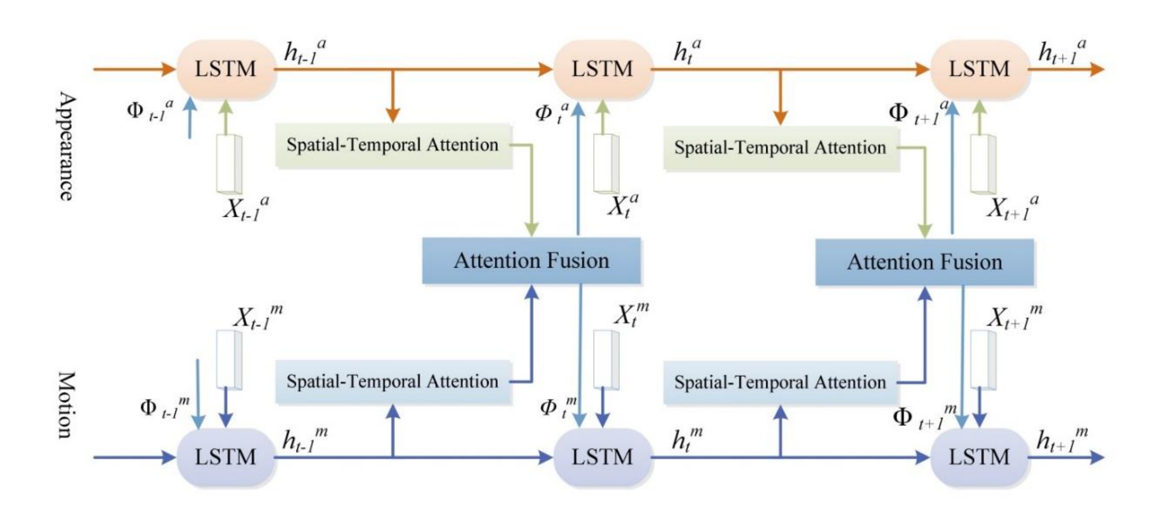


图 3 添加时空注意力机制的 LSTM 模型

2.6 决策融合规则

决策融合是将多个基分类器的结果，按照一定的规则融合成一个全局的结果，消除决策本身或决策之间的信息缺陷，提升全局结果的可靠性和稳定性[25]。本文的网络结构包含三个部分，一部分是在卷积神经网络的 STPP 层进行特征融合后的融合流，另两个部分是在将特征融合后仍然保留时间流和空间流之后的结构，并且加入注意力机制，形成两个以捕获复杂的时空线索对融合流的识别结果进行修正的分支。因此对于数据集的每一个分割数据集，网络结构最后都有三个基分类器的识别结果。对于这三个基分类器得到的分类结果，采用决策融合的方式得到最终的分类输出。

设 $C_j(\mathrm{X})(j=1,2,\cdots,N)$ 为最终融合分类结果，则融合规则可用公式表述如下，

$$C_j(X)=\arg\max_{j=1,2,\cdots,N}\left(\sum_{i=1}^{3}\omega_j\cdot\ln\left(p(X_i|c_j)\right)\right) \tag{19}$$

其中：X_i 为第 i 个基分类器的源特征，$i=1,2,3$，$\ln(p(X_i|c_j))$ 为分类器选取每一类别（c_j）分类时产生的可信度，$j=1,2,\cdots,N$，ω_j 表示融合分类的权值，其值为每个基分类器的分类精度，即单体分类精度。

于是，通过时域、空域和融合域的基分类器获得源分类结果，再利用用式(19)对源分类结果进行融合，得到数据集的每一分割集的识别分类结果。

3 实验分析

3.1 数据集和评估指标

本文实验的数据集来源于两个公开的视频动作识别数据集：UCF101 和 HMDB51。UCF101 包含 13320 个视频段，共 101 个动作类别，涵盖了较大范围的人体动作，比如化妆、打字、吹头发、骑马、跳高等。该数据集的大多数视频是在无约束的真实环境下拍摄的，因此视频存在像素低、受到如光照、遮挡等环境因素影响的问题。HMDB51 包含 6766 个视频段，共 51 个动作类别。该数据集的视频大多来源于电影剪辑片段，像素较低，主要的动作类别有亲吻、拥抱、骑马和开枪等。

实验中，本文将两个数据集分割成三份，均对其进行训练和测试。其中，每份 UCF101 的视频序列有 9500 段，HMDB51 有 3700 个视频段。由于本文网络模型有时间流、空间流和融合流三个部分，对于数据集的每一个分割集，本文对上述三个基分类器的结果进行线性加权融合得到分割集的最终动作识别准确率。线性加权融合的识别置信度权值为自适应动态权值，由测试集在基分类器的识别结果计算得出。在得到数据集的三个分割集的最终识别准确率后，对三个分割集的结果进行线性加权平均，从而得到该数据集的最终识别准确率。本文将数据集的最终识别准确率值作为人体动作识别模型的评估指标。

3.2 预训练

与图像数据集相比，人体动作识别的数据集相对较小，而对于较深的神经网络，数据集较小很容易使得网络陷入过拟合现象，因此对本文模型进行预训练。对于输入为 RGB 图像的空间域网络，直接采用图片数据库 ImageNet[26]对其进行预训练。输入的训练图片为经过数据增强扩大后的训练集，然后对其进行随机位置裁剪，并将输入大小调整为 224×224。对于输入为光流数据的时间流网络，采用从 TL-V1[27] 中提取到的动作视频光流数据，为保证和 RGB 数据同区间，再通过线性变换将光流数据离散到 $[0,255]$ 的区间上。然后对预训练空间流网络的第一层的滤波器在通道中做平均运算，将平均后的数据复制 20 次后作为时间网络的初始化数值。

3.3 实验结果与分析

在 Linux 系统搭建的 TensorFlow 平台下进行实验。深度神经网络容易陷入过拟合现象，因此本文将模型中空间流和时间流 dropout 层的丢失率分别设置为 0.7 和 0.8。空间域初始的学习率设置为 10^{-3}，在迭代 15000 次后设置为 10^{-4}，在迭代 30000 次后停止训练。时间域初始的学习率设置为 3×10^{-3}，在迭代第 20000 次后每 20000 次学习率缩小为原来的 1/10，最大迭代次数为 80000 次。

通过本文模型的第一模块来分别提取视频序列的运动特征和外观特征。考虑到 STPP 层不同池化水平对动作识别任务有不同的影响，于是设计不同池化水平的对比实验，该实验结果

来源于仅对第一模块的双流三维卷积神经网络进行训练及测试的动作识别准确率。本文考虑 STPP 两种池化水平：$\{2\times2\times1,1\times1\times1\}$ 和 $\{4\times4\times1,2\times2\times1,1\times1\times1\}$，分别描述为 STPP-1 和 STPP-2，然后在 UCF101 数据集第一分割视频序列（split1）上进行实验。由表 1 可知，当 STPP 池化水平为 $\{4\times4\times1,2\times2\times1,1\times1\times1\}$ 时，动作识别准确率均优于 STPP-1 和最大池化，因此在后续实验中，STPP 的池化水平都设置为此标准。由表 1 还可以看出，相同网络结构下，时间域的识别率高与空间域的识别率，这说明运动信息比外观信息更能表达人体动作信息。

表 1　STPP 层不同池化水平下的动作识别准确率比较

池化标准	空间域(%)	时间域(%)
Max pooling	82.76%	85.78%
STPP-1	82.18%	87.26%
STPP-2	85.74%	89.91%

表 2 展示了在本文模型第三模块的 LSTM 模型中使用时空注意力机制与否的动作识别率结果，该识别率结果由两个数据集的三个分割集的识别率结果加权平均得到。由表 2 可以看出，在时间域和空间域上使用添加注意力机制的 LSTM 模型比不使用注意力机制的动作识别准确率高，该实验也证明添加时空注意力机制的 LSTM 模型对人体动作识别任务更有效。

表 2　LSTM 模型使用注意力机制与否的动作识别准确率比较

注意力机制	域	UCF101 (%)	HMDB51 (%)
不使用	空间域	89.73%	67.95%
	时间域	91.02%	68.13%
使用	空间域	92.52%	68.16%
	时间域	93.57%	70.52%

结合时空金字塔和注意力机制的深度神经网络模型对人体动作识别任务的识别准确率如表 3 所示。对于数据集的每一个分割集的识别准确率，均是利用决策级融合的方式对上述模型中时间域、空间域和融合域的基分类器结果进行线性加权融合得到。再对三个分割集的结果线性加权平均得到相应数据集的最终动作识别准确率。

表 3　本文模型的人体动作识别准确率

分割数据集	UCF101 (%)	HMDB51 (%)
Split1	93.95%	69.16%
Split2	94.67%	71.08%
Split3	94.13%	70.86%
线性平均	94.21%	70.50%

将本文方法和近几年动作识别领域比较典型的深度学习方法或网络模型分别在 UCF101 和 HMDB51 这两个数据集上的识别准确率进行对比。这些方法分别是文献[6]提出的双流卷积网络模型（Two-stream convolutional network）；文献[20]提出的 C3D 网络模型（3D Convolutional Networks），该模型训练了更深的 CNN 网络；文献[24]提出的时空融合网络，其网络结构是双流 VGG 模型；以及文献[28]在文献[24]基础上提出的多层金字塔融合模型。从表 4 可以看出，本文提出的方法相较于近几年的经典算法更能精确地识别视频序列中的人体动作。

表 4　不同方法在 UCF101 和 HMDB51 数据集上的动作识别准确率

方法	Year	UCF101 (%)	HMDB51 (%)
Two-steam[6]	2014	88.0	59.4
C3D[20]	2015	85.2	-
Two-steam VGG[24]	2016	92.5	65.4
SPN-VGG-16[28]	2017	93.2	66.1
本文方法		94.2	70.5

4　结束语

目前基于深度学习的方法已经广泛应用到模式识别等各个领域的研究组中，对于人体动作识别任务，本文提出了改进后的结合时空金字塔和注意力机制的深度神经网络模型，构建了时空双流深度神经网络架构。将本文模型先在 ImageNet 上进行预训练和微调，然后应用到 UCF101 和 HMDB51 数据集上，通过融合时空网络与融合流最后分别取得了 94.2%和 70.5%的识别准确率。实验表明本文提出的改进深度学习模型对数据集中人体动作能够有效识别，但是对于将其应用到实际商业应用中还有一定的距离。因此，今后可以对环境因素影响较大或噪声较多的视频进行鲁棒性的算法研究。

参考文献：

[1] Mur O, Frigola M, Casals A. Modelling daily actions through hand-based spatio-temporal features [C]// Proc of Imternational Conference on Advanced Robotics. Piscataway, NJ: IEEE Press, 2015: 478-483.

[2] Liu Fang, Xu Xiangmin, Qiu Shuoyang, *et al.* Simple to complex transfer learning for action recognition [J]. IEEE Trans on Image Processing, 2016, 25 (2): 949-960.

[3] Uddin A, Joolee J B, Alam A, *et al.* Human action recognition using adaptive local motion descriptor in Spark [J]. IEEE Access, 2017, 5: 21157-21167.

[4] 黄晓晖，董超俊. 一种基于深度图去噪与时空特征提取的动作识别方法 [J]. 现代工业经济和信息化, 2017, 2017 (5): 64-68. (Huang Xiaohui, Dong Chaojun. The depth map denoising and spatio-temporal feature extraction for human action recognition [J]. Modern Industrial Economy and Informationization, 2017, 2017 (5): 64-68.)

[5] 张杰，吴剑章，汤嘉立，等. 基于时空图像分割和交互区域检测的人体动作识别方法 [J]. 计算机应用研究, 2017, 34 (1): 302-305. (Zhang Jie, Wu Jiangzhang, Tang Jiali, *et al.* Human action recognition method based on spatio-temporal image segmentation and interactive area detection [J]. Application Research of Computers, 2017, 34 (1): 302-305.)

[6] Simonyan K, Zisserman A. Two-Stream convolutional networks for action recognition in videos [J]. Advances in Neural Information Processing Systems, 2014, 1 (4): 568-576.

[7] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015-04-10) . https://arxiv. org/abs/1409. 1556.

[8] Chéron G, Laptev I, Schmid C. P-CNN: pose-based CNN features for action recognition [C]// Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2015: 3218-3226.

[9] Srivastava N, Mansimov E, Salakhutdinov R. Unsupervised learning of video representations using LSTMs [C]// Proc of the 32nd International Conference on Machine Learning. [S. l.] : International Machine Learning Society (IMLS) , 2015: 843-852.

[10] Krishnan K, Prabhu N, Babu R V. ARRNET: Action recognition through recurrent neural networks [C]// Proc of International Conference on Signal Processing and Communications. 2016: 1-5.

[11] Wang Limin, Xiong Yuanjun, Wang Zhe, *et al.* Temporal segment networks: Towards good practices for deep action recognition [C]// Proc of European Conference on Computer Vision. Berlin: Springer, 2016: 20-36.

[12] Kar A, Rai N, Sikka K, *et al.* AdaScan: adaptive scan pooling in deep convolutional neural networks for human action recognition in videos [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 5699-5708.

[13] Du Wenbin, Wang Yali, Qiao Yu. Recurrent spatial-temporal attention network for action recognition in videos [J]. IEEE Trans on Image Processing, 2017, 27 (3): 1347-1360.

[14] Ji Shuiwang, Xu Wei, Yang Ming, *et al.* 3D convolutional neural networks for human action recognition [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2013, 35 (1): 221-231.

[15] Veeriah V, Zhuang Naifan, Qi Guojun. Differential recurrent neural networks for action recognition [C]// Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2015: 4041-4049.

[16] Ordóñez F J, Roggen D. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition [J]. Sensors, 2016, 16 (1): 115-140.

[17] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9 (8): 1735-1780.

[18] 陈胜娣，魏维，何冰倩，等. 基于改进的深度卷积神经网络的人体动作识别方法 [J/OL]. 计算机应用研究, 2019, 36 (4) . (2018-02-09) [2018-08-23]. http://www. arocmag. com/article/02-2019-04-054. html. (Chen Shengdi, Wei Wei, He Bingqian, *et al.* Human action recognition base on improved deep convolutional neural networks [J]. Application Research of Computers, 2019, 36 (4) . (2018-02-09) [2018-08-23]. http://www. arocmag. com/article/02-2019-04-054. html.)

[19] Karpathy A, Toderici G, Shetty S, *et al.* Large-scale video classification with convolutional neural networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 1725-1732.

[20] Du Tran, Bourdev L, Rob Fergus, *et al.* learning spatiotemporal features with 3D convolutional networks [C]// Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2015: 4489-4497.

[21] Sun Lin, Jia Kui, Yeung D Y, *et al.* Human action recognition using factorized spatio-Temporal convolutional networks [C]// Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2015: 4597-4605.

[22] Liu Li, Shao Lin, Li Xuelong, *et al.* Learning spatio-temporal representations for action recognition: a genetic programming approach [J]. IEEE Trans on Cybernetics, 2016, 46 (1): 158-170.

[23] Wang Miao, Sun Jifeng, Yu Jialin, *et al.* Human action recognition based on feature level fusion and random projection [C]// Proc of the 5th International Conference on Computer Science and Network Technology. Piscataway, NJ: IEEE Press, 2016: 767-770.

[24] Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 1933-1941.

[25] 张文宇. 基于证据理论的无线传感器网络决策融合算法研究 [D]. 北京：北京交通大学, 2016. (Zhang Wenyu. Research on belif function based decision fusion for wireless sensor networks [D]. Beijing: Bejing Jiaotong University, 2016.)

[26] Deng Jia, Dong Wei, Socher R, *et al.* ImageNet: A large-scale hierarchical image database [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2009: 248-255.

[27] Pérez J S. TV-L1 optical flow estimation [J]. Image Processing on Line, 2013, 2 (4): 137-150.

[28] Yu Yunbo, Long Mingsheng, Wang Jianmin, *et al.* Spatiotemporal pyramid network for video action recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 2097-2106.